Distributed Computation Resources for the Earth System Grid Federation (ESGF)

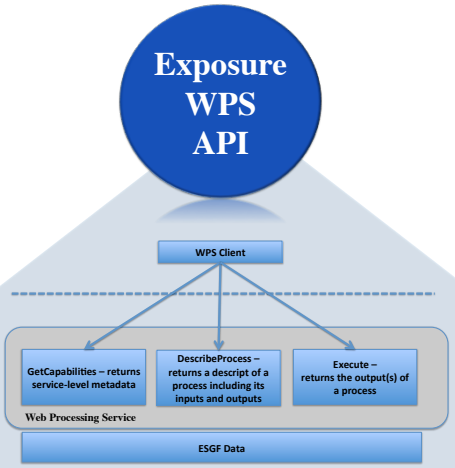
Daniel Q. Duffy[1], Charles Doutriaux[2], Dean Williams[2]
[1]NASA Center for Climate Simulation (NCCS), NASA Goddard Space Flight Center (GSFC); [2]Lawrence Livermore National Laboratory (LLNL)

Data Analytics for ESGF

The Intergovernmental Panel on Climate Change (IPCC), prompted by the United Nations General Assembly, has published a series of papers in their Fifth Assessment Report (AR5) on processes, impacts, and mitigations of climate change in 2013 and 2014. The science used in these reports was generated by an international group of domain experts using highly complex models to simulate the Earth's climate over long periods of time. The resulting total data of approximately five petabytes are stored in a distributed data grid known as the Earth System Grid Federation (ESGF). Through the ESGF, consumers of the data can find and download data with limited capabilities for server-side processing.

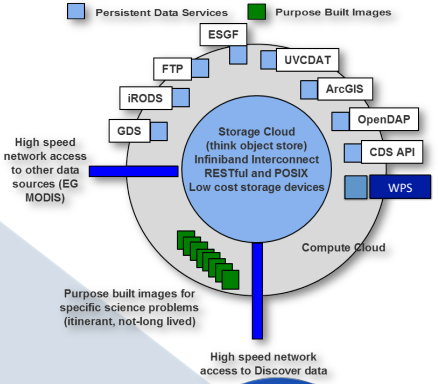
The Sixth Assessment Report (AR6) is already in the planning stages and is estimated to create significantly more data than AR5. A major paradigm shift from downloading data to perform analytics must evolve to moving the analysis routines to the data. In preparation, the ESGF has started a Compute Working Team (CWT) to create solutions that allow users to perform distributed, high-performance data analytics on the AR6 data. The team has started to design and develop a general Application Programming Interface (API) to enable highly parallel, server-side processing throughout the ESGF data grid federation.

Over the next year (2015), this team will be working to create a Web Processing Service (WPS) API specification and a reference implementation.



High-Performance Compute and Storage Fabric

Storage-proximal analytics – The API exposes methods by which the analytics are transferred to where the data resides. The data is stored within a high-performance compute and storage fabric using a combination of high-performance computing (HPC) and cloud computing capabilities. The same data may be accessible through the ESGF along with a number of other web services.



Data

Compute Storage

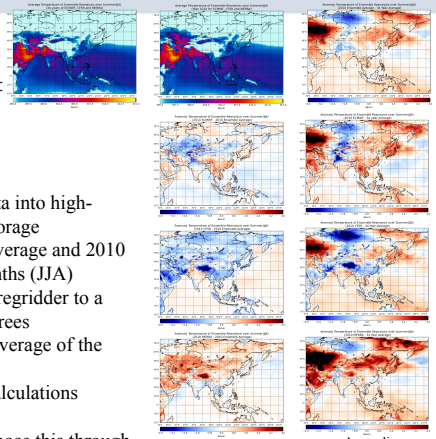
Use Case – Anomaly

Representative Use Case: Multi-Model Averaging
Generate an average of a variable over many models across federated data.
Specify the temporal and spatial extent over which the averaging will be done.

These figures illustrate anomalies for surface temperature over three different reanalysis datasets:
- ECMWF
- MERRA
- CFSR

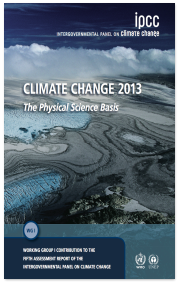
Process
- Download and ingest data into high-performance compute/storage
- Generate both 34-year average and 2010 average for summer months (JJA)
- Regrid using the ESMF regridder to a common grid ¾ x ¾ degrees
- Compute the ensemble average of the regridded data
- Compute the anomaly calculations

The WPS service would expose this through an API to enable these calculations to be performed where the data reside!



Relevance and Collocation

The data exposed must have a high relevance to the scientific community. In addition, the data must be both physically and logically collocated (i.e., federated).



The data stored across the ESGF is a very large and highly complex set of simulation and observation data.

Estimated Data Growth of ESGF
2012 AR5 – 2 to 5 petabytes
2017 AR6 – 25 to 50 petabytes (5x to 10x)
2022 AR7 – 100 to 1,000 petabytes

ESGF Compute Working Team

The Earth Systems Grid Compute Working Team is an international group consisting of members from ESGF data grid sites.

Co-Chairs
- Daniel Duffy (NASA/GSFC) daniel.q.duffy@nasa.gov
- Charles Doutriaux (DOE/LLNL) doutriaux1@llnl.gov

References:
- Schnase, et al., "MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service," *Computers, Environment, and Urban Systems* (doi:10.1016/j.compenvurbsys.2013.12.003)
- http://en.wikipedia.org/wiki/Web_Processing_Service